16 Analyzing, synthesizing and converting speech prosody

～Generative modeling of voice fundamental frequency contours～

Abstract

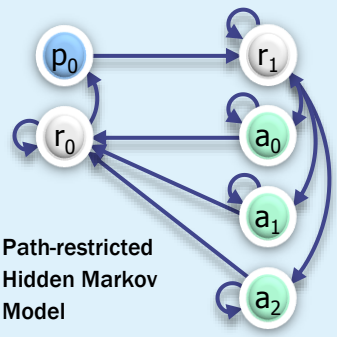
Linear Predictive Coding (LPC), proposed in the 60s, has established the modern speech analysis/synthesis framework and has opened the door of today's mobile and VoIP communication technology. While LPC has realized the analysis/synthesis framework focusing on the 'phonemic' factor of speech, the aim of this work is to develop a new analysis/synthesis framework focusing on the 'prosodic' factor. Although a well-founded physical model for vocal fold vibration was proposed in the 60s by Fujisaki (known as the "Fujisaki model"), how to estimate the underlying parameters has long been a difficult task. We have developed a stochastic counterpart of the Fujisaki model, which made it possible to apply powerful statistical inference techniques to accurately estimate the underlying parameters. This model has a high potential to be developed into a next-generation module for Text-to-Speech, speech analysis, synthesis and conversion systems.

What is fundamenal frequency (F0) contour?

- An acoustic correlate that plays an important role in conveying non-linguistic information such as the identity, intention, attitude and mood of the speaker

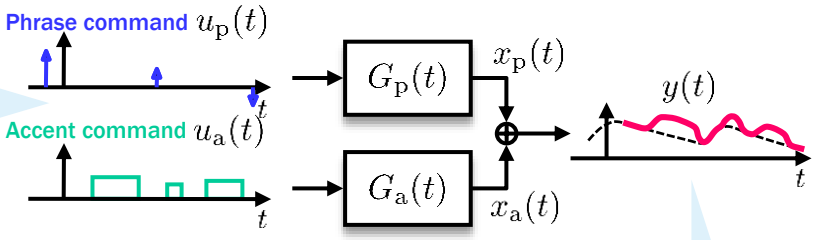
The aim of this work

- Develop stochastic counterpart of the Fujisaki model → Elegant parameter estimation framework
- Realize an analysis and synthesis framework of speech prosody and natural-sounding Text-to-Speech synthesis



Path-restricted Hidden Markov Model

Physical model for F0 generation process (Fujisaki model)



For $t = 1, \ldots, T$:

$$s_t | s_{t-1} \sim \pi_{s_{t-1}, s_t} \quad \text{(state sequence)}$$

$$\begin{bmatrix} u_{\mathrm{p}}[t] \\ u_{\mathrm{a}}[t] \end{bmatrix} | s_t \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\mathrm{p},s_t} \\ \mu_{\mathrm{a},s_t} \end{bmatrix}, \begin{bmatrix} \sigma_{\mathrm{p}}^2 & 0 \\ 0 & \sigma_{\mathrm{a}}^2 \end{bmatrix} \right)$$

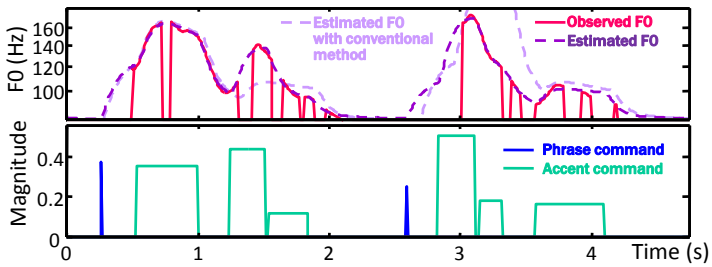
Change of variables

Generative model of F0 contours

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\begin{cases} \boldsymbol{\mu} = \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\mu}_{\mathrm{a}} + \mu_{\mathrm{b}}\boldsymbol{1} \\ \boldsymbol{\Sigma} = \sigma_{\mathrm{p}}^2 \boldsymbol{G}_{\mathrm{p}}\boldsymbol{G}_{\mathrm{p}}^{\mathsf{T}} + \sigma_{\mathrm{a}}^2 \boldsymbol{G}_{\mathrm{a}}\boldsymbol{G}_{\mathrm{a}}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathrm{b}} \end{cases}$$

Estimated phrase and accent commands with proposed method



Audio demo

What will become possible?

- Manually changing phrase and accent commands allows to convert intonation as desired while keeping naturalness of speech
- Predicting phrase and accent commands from text inputs allows to synthesize speech with natural-sounding F0 contours

Related works

[1] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," IEEE/ACM Transactions on Audio, Speech and Language Processing, to appear, 2015.

[2] K. Kadowaki, T. Ishihara, N. Hojo, and H. Kameoka, "Speech prosody generation for text-to-speech synthesis based on generative model of F0 contours," in Proc. The 15th Annual Conference of the International Speech Communication Association (Interspeech 2014), pp. 2322-2326, Sep. 2014.

Contact

Hirokazu Kameoka Recognition Research Group, Media Information Laboratory
E-mail : kameoka.hirokazu(at)lab.ntt.co.jp